Automatically Measuring the Quality of User Generated Content in Forums

Kevin Chai, Chen Wu, Vidyasagar Potdar, and Pedram Hayati

Digital Ecosystems and Business Institute, Curtin University
International Centre for Radio Astronomy Research,
University of Western Australia
{k.chai,p.hayati}@curtin.edu.au, chen.wu@uwa.edu.au,
vidysagar.potdar@cbs.curtin.edu.au

Abstract. The amount of user generated content on the Web is growing and identifying high quality content in a timely manner has become a problem. Many forums rely on its users to manually rate content quality but this often results in gathering insufficient rating. Automated quality assessment models have largely evaluated linguistic features but these techniques are less adaptive for the diverse writing styles and terminologies used by different forum communities. Therefore, we propose a novel model that evaluates content, usage, reputation, temporal and structural features of user generated content to address these limitations. We employed a rule learner, a fuzzy classifier and Support Vector Machines to validate our model on three operational forums. Our model outperformed the existing models in our experiments and we verified that our performance improvements were statistically significant.

Keywords: content quality assessment, user generated content, forums.

1 Introduction

Forums websites allows people to engage in online discussions. There are millions of forums on the Web and each forum can host large volumes of User Generated Content (UGC). However, forum users are being overwhelmed with excessive amounts of UGC and it is becoming more difficult to identify high quality content in a timely manner. Currently, many forums and Web 2.0 websites rely on its users to manually rate the quality of content to handle this problem [5]. However, there are a number of problems with relying solely on user ratings. Firstly, rating is voluntary so a large percentage of content often receives a lack of rating [22,19]. Secondly, users may not have sufficient knowledge and expertise to provide accurate ratings [17]. Lastly, reliance on manual user ratings becomes an ongoing problem if UGC is created at a faster speed than which it can be sufficiently rated [4]. Therefore, the objective of this paper is to propose a novel model that automatically measures the quality of UGC in forums. More specifically, the contributions of this paper are to:

D. Wang and M. Reynolds (Eds.): AI 2011, LNAI 7106, pp. 51–60, 2011.

- Present a model that evaluates content, usage, reputation, temporal and structural features for assessing forum post quality.
- Validate our model against three operational forums using supervised machine learning techniques and compare its performance against existing models in the literature.

2 Problem Definition

We formally define the problem of measuring the quality of forum posts as a multi-class classification problem. The forum dataset is described by a set of posts $P = \{p_1, p_2, ..., p_i, ..., p_{|P|}\}$ and a set of post quality classes $C = \{c_1 = low, c_2 = medium, c_3 = high\}$ where p_i is the i-th post in P. Furthermore, posts are represented as a set of content quality features $F = \{f_1, f_2, ..., f_j, ..., f_{|F|}\}$ in our model as defined for p_i in 1.

$$p_i = \{f_1^i, f_2^i, ..., f_j^i, ..., f_{|F|}^i\} \tag{1}$$

$\phi(p_i, c_k)$ is a Boolean function that is used to determine whether p_i belongs to c_k where $k = \{1, 2, 3\}$ as defined in 2.

$$\phi(p_i, c_k) : P \times C \rightarrow \{True, False\} \tag{2}$$

The task of performing automated post quality classification is to evaluate this function for all posts in a given forum dataset.

3 UGCQ Assessment Model

In recent work [4] we proposed a model that measures the quality of forum posts based upon its usage within a forum community. We extend this work by proposing a UGC Quality (UGCQ) model that evaluates content, usage, reputation, structural and temporal features for quality assessment.

Content features represent intrinsic information about the forum post such as features related to its textual content. Usage features represent the popularity of postings and usage data is obtained using the post usage tracking framework developed in our previous work [4]. Usage features evaluate view counts, dwell time as well as mouse and keyboard interactions between users and posts.

Reputation features evaluate the activeness, accountability and authority of post authors to gauge their overall reputation for quality assessment. Temporal features represent time-based characteristics of postings and evaluate the timeliness of when a forum post is created and edited. Structural features evaluate the position and visibility of postings within a forum thread.

As a result, we propose 46 post quality features based on these categories in the UGCQ model as presented in Table 1. An in-depth explanation of each feature and how it is measured is provided in [3].

Table 1. UGCQ Model Features

ID	Name
	Content
f_1	Word count
f_2	Unique word count
f_3	Ratio word count to average word count in thread
f_4	Quoted word count
f_5	Original word count
f_6	Ratio original word count to word count
f_7	Formatting tag count
f_8	Ratio formatting tag count to formatting tag count in thread
f_9	Hyperlink count
f_{10}	External hyperlink count
f_{11}	Internal hyperlink count
f_{12}	Ratio hyperlink count to hyperlink count in thread
f_{13}	Attachment count
f_{14}	Ratio attachment count to attachment count in the thread
f_{15}	Attachment download count
f_{16}	Ratio attachment download count to thread downloads
f_{17}	Post edit count
f_{18}	Post reported count
f_{19}	Is post created by thread author
	Usage
f_{20}	Post view count
f_{21}	Distinct user view count
f_{22}	Distinct users that revisit in different sessions count
f_{23}	Total dwell time
f_{24}	Average dwell time
f_{25}	Text selection count
f_{26}	Total number of characters selected
f_{27}	Average number of characters selected
f_{28}	Text copy count
f_{29}	Total number of characters copied
f_{30}	Average number of characters copied
	Reputation
f_{31}	First name, last name and location provided
f_{32}	E-mail displayed to public
f_{33}	Website URL provided
f_{34}	Membership group (member, moderator, administrator)
f_{35}	Number of posts created by user
f_{36}	Membership age
	Temporal
f_{37}	Age
f_{38}	Post edit time difference
f_{39}	Previous post time difference
f_{40}	Previous post time difference to thread average difference
f_{41}	Following post time difference
	Structural
f_{42}	Is first post
f_{43}	Is displayed on first thread page
f_{44}	Is last post
f_{45}	Is displayed on last thread page
f_{46}	Thread position to thread post count

4 Experiment

4.1 Datasets

We obtain three forum datasets for evaluating the performance of the UGCQ model. Firstly, data from `http://remnantsguild.com/` was collected from the July 21, 2009 to October 16 2009. Secondly, data from `http://nabble.com/` [22] and `http://slashdot.org/` [21] are obtained for experimentation. The Nabble dataset contains data from April 1, 2002 to July 24, 2006. The Slashdot dataset contains posts created from September 10, 2007 to September 24, 2007. Details of the datasets are displayed in Table 2.

Table 2. Forum Datasets

	Remnantsguild	**Nabble**	**Slashdot**
Users	54	1,832	3,893
Topics	114	2,956	191
Rated posts	531	4,291	7,847
Low quality posts	288 (54%)	2,037 (48%)	4,026 (51%)
Medium quality posts	166 (31%)	515 (12%)	2,693 (34%)
High quality posts	77 (15%)	1,739 (40%)	1,128 (15%)

Hsu *et al.* (2003) [12] showed that they could improve the performance of their Support Vector Machines (SVM) classifier by performing data normalisation. Therefore, we adopt a min-max data normalisation approach to scale feature values to a range of [0, 1] to avoid features in larger numeric ranges from dominating those in smaller ranges. Additionally, classifier performance can be improved when continuous features are discretised into ranked intervals [8]. Therefore, we use the Fayyad & Iranis Minimum Description Length method [9] for data discretisation. The datasets are split into complementary training and test sets using 10 fold cross-validation in our experiments.

4.2 Feature Selection

A number of features in the UGCQ model could not be evaluated for the Nabble and Slashdot datasets due to missing data. For example, usage data was not collected from Nabble and Slashdot because the datasets were provided to us. We had collected usage data from the Remnantsguild forum with our post usage tracking framework we proposed in [4]. As a result, the set of features evaluated for each dataset is (refer to Table 1 for feature names):

- **Remnantsguild:** 46 features $\{f_1\text{-}f_{46}\}$
- **Nabble:** 24 features $\{f_1\text{-}f_{12}, f_{19}, f_{35}, f_{37}\text{-}f_{46}\}$ with 22 features missing
- **Slashdot:** 23 features $\{f_1\text{-}f_{12}, f_{19}, f_{35}, f_{37}, f_{39}\text{-}f_{46}\}$ with 23 features missing

We perform feature selection using a sequential forward selection approach. The purpose of conducting feature selection is to identify the set of most important and relevant features for classifying the quality of forum posts. Waikato Environment for Knowledge Analysis (WEKA) [11] is a data mining tool that we use to perform feature selection and classification in our experiments. The selected feature sets generated for each forum dataset are:

- **Remnantsguild:** 8 features $\{f_1\text{-}f_3, f_5, f_8, f_9, f_{24}, f_{38}\}$
- **Nabble:** 4 features $\{f_5, f_{12}, f_{35}, f_{37}\}$
- **Slashdot:** 4 features $\{f_{35}, f_{39}, f_{45}, f_{46}\}$.

4.3 Performance Evaluation

We use the classification accuracy and Matthews Correlation Coefficient (MCC) to evaluate the performance of our model and existing models in the literature. MCC is considered one of the best for evaluating classifier performance on the imbalanced data [2] as in our experiment (See Table 2). This metric provides a correlation value between -1 to 1 where -1 represents perfect inverse prediction, 0 represents random prediction and 1 represents perfect prediction.

A MCC value is calculated from a classifiers confusion matrix for each quality class (i.e. low, medium and high). The MCC performance measure is defined in 3 where TP = true positives, TN = true negatives, FP = false positives and FN = false negatives.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

4.4 Post Quality Classification

We developed the UGCQ model into a working prototype and implemented the models proposed by Weimer & Gurevych (2007) [22] and Wanas *et al.* (2008) [21]. We classify the quality of forum posts from each dataset using WEKA [11]. More specifically, we use WEKA's implementation of the Sequential Minimal Optimisation (SMO) algorithm [18] for SVM, the rule based learner JRIP which is based on Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [6] as well as the fuzzy rule learner FURIA [13].

We perform a number of classification experiments so we introduce a naming scheme to label each experiment in the form of [model_name]_[classifier]_[encoding]_[selection] and the values of each of these fields is displayed in Table 3. For example, the experiment with the UGCQ framework, FURIA algorithm, normalisation and feature selection is labelled as UGCQ_FURIA_N_FS.

Table 3. Forum Post Classification Experiment Labels

Field	Values	Comments
model_name	UGCQ, Weimer, Wanas, Baseline	Baseline refers to the majority class classifier
classifier	JRIP, FURIA, SVM	Repeated Incremental Pruning to Produce Error Reduction (RIPPER) variant algorithm, Fuzzy Unordered Rule Induction Algorithm, Support Vector Machines
encoding	N, D	Normalisation, discrestiation
selection	FS	Feature selection

5 Results

The experimental results obtained from the Remnantsguild, Nabble and Slashdot datasets are presented in Table 4. The UGCQ model using the JRIP on the normalised Remnantsguild dataset achieved the best results with 68.55% accuracy and an average MCC value of 0.45. The Weimer model also achieved 68.55% accuracy but with a lower average MCC of 0.43 while the Wanas model achieved 63.84% with an average MCC value of 0.30. All CQA models outperformed the majority class baseline of 54.24% for Remnantsguild.

The UGCQ model using SVM on the discretised Nabble dataset achieved the best results with 69.98% accuracy and an average MCC value of 0.40. Additionally, the Weimer model achieved 65.07% accuracy with an average MCC of 0.28 while the Wanas model achieved 58.24% with an average MCC value of 0.19. All CQA models outperformed the majority class baseline of 47.47% for Nabble.

The UGCQ model using SVM on the normalised Slashdot dataset achieved the best accuracy of 53.94% accuracy but with an average MCC value of 0.12. The FURIA algorithm on the discretised dataset however achieved the highest average MCC value of 0.15 but with a lower accuracy of 51.20%. Additionally, the Weimer model achieved 51.29% accuracy with an average MCC of 0 while the Wanas model achieved 51.31% with an average MCC value of 0.01. The UGCQ model using SVM on the normalised dataset slightly outperformed the majority class baseline of 51.31% while the Weimer model under performed and the Wanas model achieved equivalent performance to the baseline for Slashdot.

5.1 Friedman Test

Demšar (2006) [7] surveyed papers published from the International Conference of Machine Learning in 1999 to 2003 and discovered that the majority of authors did not statistically verify whether their classifier(s) produced significant performance improvements.Therefore, a number of suitable statistical tests were

Table 4. Ranking Comparison of Classifiers over all Datasets

Classifier	Remnantsguild	Nabble	Slashdot	Rank_{avg}
UGCQ_SVM_D	66.29% (4)	69.98% (1)	52.66% (2.5)	2.5
UGCQ_JRIP_N	68.55% (1.5)	68.49% (3)	52.31% (4.5)	3
UGCQ_SVM_D_FS	65.16% (7)	68.10% (4)	52.66% (2.5)	4.5
UGCQ_JRIP_D	66.85% (3)	66.98% (7)	52.31% (4.5)	4.83
UGCQ_FURIA_N	65.35% (6)	67.78% (5)	52.03% (6)	5.67
UGCQ_JRIP_N_FS	63.47% (10)	68.93% (2)	51.56% (7)	6.33
UGCQ_SVM_N	63.65% (9)	64.18% (12)	53.94% (1)	7.33
Weimer	68.55% (1.5)	65.07% (11)	51.29% (13)	8.5
UGCQ_FURIA_D	65.72% (5)	66.25% (8)	51.20% (14)	9
UGCQ_JRIP_D_FS	61.39% (13)	66.16% (9)	51.31% (8)	10
UGCQ_FURIA _D_FS	62.90% (11)	66.05% (10)	51.31% (10.5)	10.5
Wanas	63.84% (8)	58.24% (14)	51.31% (10.5)	10.83
UGCQ_FURIA_N_FS	62.71% (12)	67.28% (6)	50.69% (15)	11
UGCQ_SVM_N_FS	59.89% (14)	63.34% (13)	51.31% (10.5)	12.5
Baseline	54.24% (15)	47.47% (15)	51.31% (10.5)	13.50

recommended based on the characteristics of a given experiment. We follow this recommendation by performing the Friedman test [10] for verifying if there is a significant statistical difference between the performance of multiple classifiers over multiple datasets.

Firstly, we rank classifiers within each dataset in terms of their classification accuracy. We use accuracy rather than the MCC average to include the baseline classifier for evaluation. The average rank for each classifier over all the datasets is presented in Table 4 in decreasing order of rank. Secondly, we evaluate the null hypothesis H_0 and alternate hypothesis H_a to determine if the average ranks of these classifiers over all datasets are significantly different:

- H_0: There is no difference in the average ranks for classifiers over the datasets.
- H_a: A difference exists in the average ranks for classifiers over the datasets.

We use statistical analysis tool, R [20] and conducted the Friedman test [10] to obtain a chi-squared χ^2 value of 24.84 with 14 degrees of freedom df and a p-value of 0.03618. The critical value of α based on the χ^2 value and df for the χ^2 distribution is 0.05. Therefore, we reject H_0 and accept H_a because 0.03618 (p-value) < 0.05 (α).

5.2 Nemenyi Test

We discovered from the Friedman test that some classifiers are significantly different to others but we do not know which specific classifiers are different. Therefore, we can use the Nemenyi test [16] to evaluate all pairs of classifiers ($\sum_{i=1}^{k-1} i$ permutations) to determine which classifiers are significantly different to each

other. The critical distance q_α for the two-tailed Nemenyi test with $\alpha = 0.05$ (significance level) and $k = 15$ (number of classifiers) is 3.391.

We first calculate the distance between the average ranks between all pairs of classifiers. The distance between the average ranks of two classifiers must be ≥ 3.391 to be considered as significant with 95% probability. 56 out of 105 significant differences were identified from the pair-wise comparisons between the classifiers.

We compare our top UGCQ classifier (UGCQ_SVM_D) along with the existing models in the literature as shown in Table 5. The number shown in parenthesis depicts the rank of the classifier over all datasets identified from Table 4. These results show that the performance of the UGCQ classifier is significantly different from these models while the differences between the Weimer and Wanas classifier and, the Wanas and Baseline classifier are not significant.

Table 5. Comparisons between UGCQ and Existing CQA Models

Classifier A	**Classifier B**	**Difference**	**Sig. (diff $\geq$ 3.391)**
UGCQ_SVM_D (1)	Weimer (8)	6.00	Yes
UGCQ_SVM_D (1)	Wanas (12)	8.30	Yes
UGCQ_SVM_D (1)	Baseline (15)	11.00	Yes
Weimer (8)	Wanas (12)	2.33	No
Weimer (8)	Baseline (15)	5.00	Yes
Wanas (12)	Baseline (15)	2.67	No

6 Discussion

Seven out of twelve UGCQ classifiers outperformed the CQA models proposed by [22], [21] over the three datasets as shown in Table 4. Additionally, we statistically verified our highest ranking UGCQ classifier (UGCQ_SVM_D) significantly outperformed these models as highlighted in Table 5.

A large number of UGCQ features were not evaluated for the Nabble and Slashdot datasets due to missing data. For example, the average dwell time was identified as an important quality feature on the Remnantsguild dataset but could not be evaluated for the other datasets. The inclusion of our missing features could further improve the performance of the UGCQ classifiers.

We calculate the average MCC low, medium and high values excluding the baseline classifier for each dataset. The results indicate that CQA models performed better in classifying low and high quality posts than medium quality. This supports our intuition of how classifiers could misclassify low and high quality posts that neighbour closely with the medium quality class and vice versa.

7 Related Work

Chai *et al.* (2009) [5] conducted a comprehensive review of 19 content quality related assessment frameworks for forums, question & answering (Q&A) websites, blogs and wikis. Additionally, Zhu *et al.* 2009 [23] proposed and validated a multidimensional framework for assessing the quality of answers in Q&A websites.

Weimer & Gurevych (2007) [22] was first to propose a model for measuring the quality of forum posts and classified posts into two quality classes (high and low) by assessing surface, lexical, syntactic, similarity and forum specific post features. This work was extended by Wanas *et al.* (2008) [21] by classifying posts into 3 quality classes (low, medium and high) and evaluated features such as relevance, originality, post component, surface and forum-specific features. Lui & Baldwin (2009) [15] evaluated bag-of-words features and features proposed by [21] on the dataset collected by [22] for classifying good and bad posts.

Agichtein *et al.* (2008) [1] evaluated usage statistics of questions and answers in Yahoo! Answers to find high quality content. Additionally, the number of times an answer was copied by users was proposed as a feature by Jeon *et al.* (2006) [14] for measuring the quality of answers in Naver! (Korean Q&A website). Our previous work, Chai *et al.* (2010) [4] extended these ideas to track how users interact with forum posts to predict its quality.

We gained a number of insights from these related studies to propose our UGCQ model that measures the content, usage, reputation, temporal and structural features of UGC for quality assessment. We provide a detailed review of the related work in the area of content quality assessment in Chai (2011) [3].

8 Conclusion

We have proposed the UGCQ model that evaluates the content, usage, reptuation, temporal and structural features of forum UGC for quality assessment. We implemented our model into a prototype and validated its performance on the Remnantsguild, Nabble and Slashdot forums. Additionally, we implemented two existing models in the literature for performance comparison with the UGCQ model. We discovered that our model outperformed the existing models in the literature over all forum datasets and the performance increase was statistically significantly.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding High-Quality content in social media. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM), pp. 183–194 (2008)
2. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics 16(5), 412–424 (2000)
3. Chai, K.: A Machine Learning-based Approach for Automated Quality Assessment of User Generated Content in Web Forums. Ph.D. thesis, Curtin University (2011)
4. Chai, K., Hayati, P., Potdar, V., Wu, C., Talevski, A.: Assessing post usage for measuring the quality of forum posts. In: Proceedings of the 4th IEEE International Conference on Digital Ecosystems and Technologies, DEST (2010)

5. Chai, K., Potdar, V., Dillon, T.: Content Quality Assessment Related Frameworks for Social Media. In: Gervasi, O., Taniar, D., Murgante, B., Laganà, A., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2009. LNCS, vol. 5593, pp. 800–814. Springer, Heidelberg (2009)
6. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning, p. 115 (1995)
7. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)
8. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th International Conference on Machine Learning, pp. 194–202 (1995)
9. Fayyad, U., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the International Joint Conference on Uncertainty in Artifical Intelligence, pp. 1022–1027 (1993)
10. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200), 675–701 (1937)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations 11(1) (2009)
12. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Tech. rep., National Taiwan University (2003), http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf
13. Hühn, J., Hüllermeier, E.: FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery 19(3), 293–319 (2009)
14. Jeon, J., Croft, W.B., Lee, J.H., Park, S.: A framework to predict the quality of answers with Non-Textual features. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 228–235 (2006)
15. Lui, M., Baldwin, T.: You are what you post: User-level features in threaded discourse. In: Proceedings of the Fourteenth Australasian Document Computing Symposium (ADCS 2009), pp. 98–105 (2009)
16. Nemenyi, P.: Distribution-free multiple comparisons. Ph.D. thesis, Princeton University (1963)
17. Nussbaum, M.E., Hartley, K., Sinatra, G.M., Reynolds, R.E., Bendixe, L.D.: Enhancing the quality of On-Line discussions. In: Paper Presented at the Annual Meeting of the American Educational Research Association (2002)
18. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in Kernel Methods Support Vector Learning 208(MSR-TR-98-14), 1–21 (1998)
19. Suryanto, M., Lim, E.P., Sun, A., Chiang, R.: Quality-Aware collaborative question answering: Methods and evaluation. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 142–151 (2009)
20. Team, R.D.C.: R: A Language and Environment for Statistical Computing. Vienna, Austria (2011), http://www.R-project.org
21. Wanas, N., El-Saban, M., Ashour, H., Ammar, W.: Automatic scoring of online discussion posts. In: Proceeding of the 2nd ACM Workshop on Information Credibility on the Web, pp. 19–26 (2008)
22. Weimer, M., Gurevych, I.: Predicting the perceived quality of web forum posts. In: Proceedings of the Conference on Recent Advances in Natural Language Processing (2007)
23. Zhu, Z., Bernhard, D., Gurevych, I.: A Multi-Dimensional model for assessing the quality of answers in social Q&A sites. Tech. rep., Ubiquitous Knowledge Processing Lab (2009)